Universidade Cruzeiro do Sul

NOVA DIDÁTTICA
EDUCAÇÃO E DESENVOLVIMENTO

Estatística Multivariada I

Regressão Linear Múltipla

Prof. Dr. Juliano van Melis

Parte II

Parte 1

Conteúdo

- Introdução a Regressão linear Múltipla
  - Pressupostos da RLM
  - Análise Gráfica dos resíduos
  - Detectar Outliers
    - Distancia de Cook
- RLM com R

Parte 2

Conteúdo

- Introdução a Regressão linear Múltipla
  - Detectar Outliers
    - Distância de Mahalanobis
    - Pontos influentes (*dffits* e *dfbeta*)
    - Resíduos studantizados
  - Indicadores de multicolinearidade
  - Fator de Inflação de Variância (VIF) e Tolerância
  - Coeficiente Explicação Ajustado
- RLM com variáveis categóricas *Dummy*
- Aplicação prática: tratamento e modelagem dos dados.
- **Avaliação**

https://xkcd.com/1838/

CORRELAÇÃO DAS VARIÁVEIS

Relembrando

Vamos agora visualizar o valor das correlações.

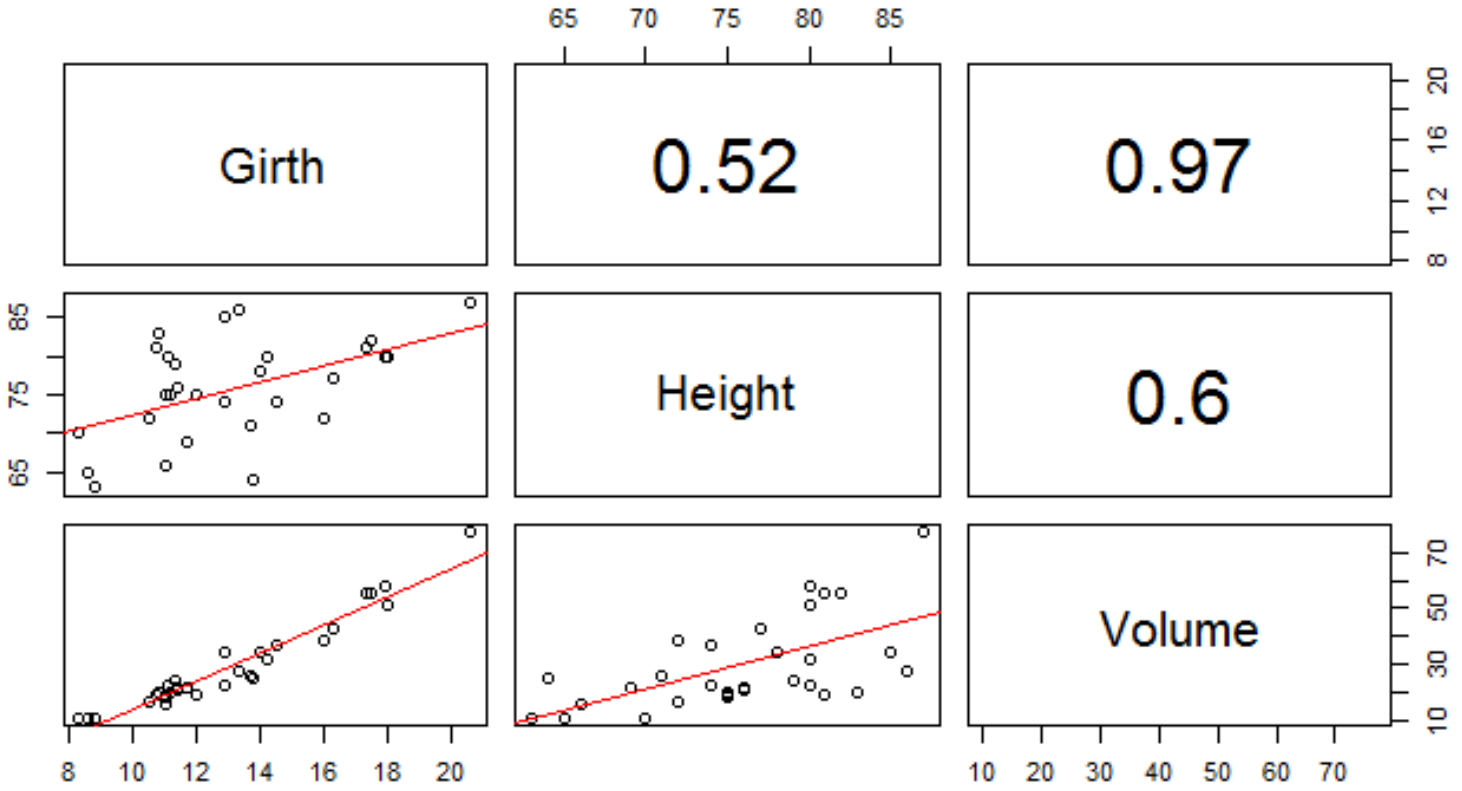
Isso pode ser feito com os argumentos `lower.panel` e `upper.panel` da seguinte forma:

```
#definindo uma função para desenhar retas de regressão:
flines<- function(x,y){
  points(x,y)
  abline(lm(y~x), col="red")
}

#definindo uma função para plotar as correlações:
fcor<- function(x,y){
  par(usr=c(0,1,0,1))
  txt<- as.character(round(cor(x,y),2))
  text(0.5, 0.5, txt, cex=3)
}
```

Relembrando

Vamos agora plotar o gráfico para essas correlações:

```
pairs(trees, lower.panel=flines, upper.panel = fcor)
```

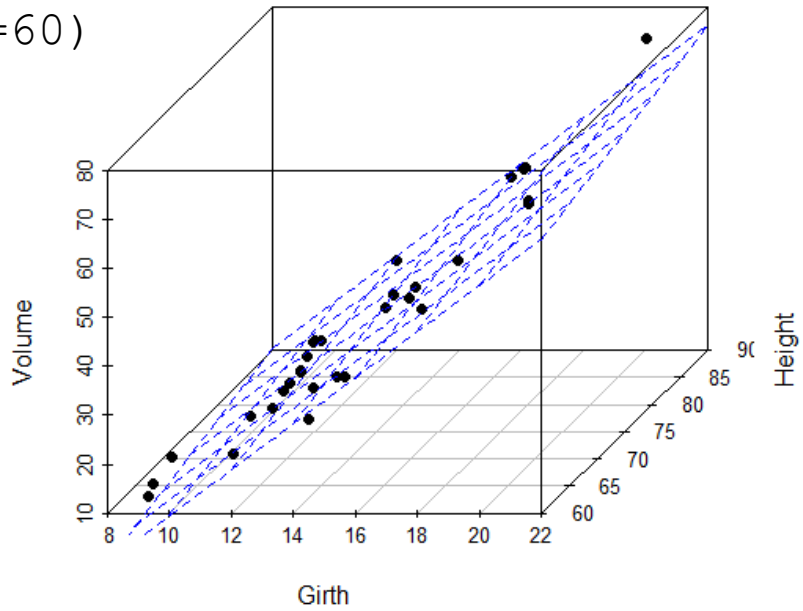


Visualização em 3D

Vamos usar o `scatterplot3d` para visualizar as três variáveis do dataset "`trees`" em um gráfico de dispersão e ainda inserir um plano de regressão.

Para isso, execute os comandos a seguir:

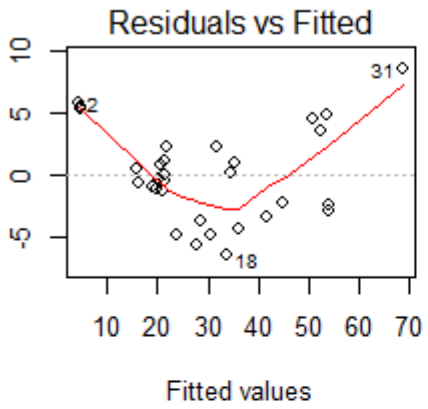
```
library(scatterplot3d)
attach(trees)
graph<- scatterplot3d(Volume ~ Girth + Height,
                      pch=16, angle=60)
fit <- lm(Volume ~ Girth + Height)
graph$plane3d(fit, col="blue")
```

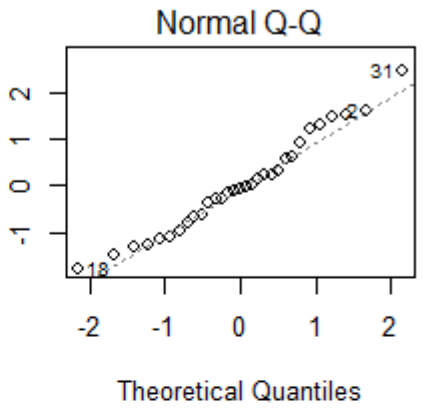


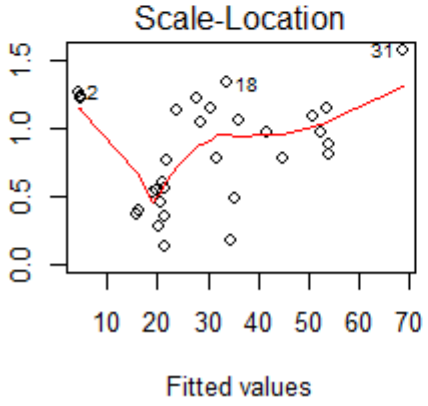
Relembrando

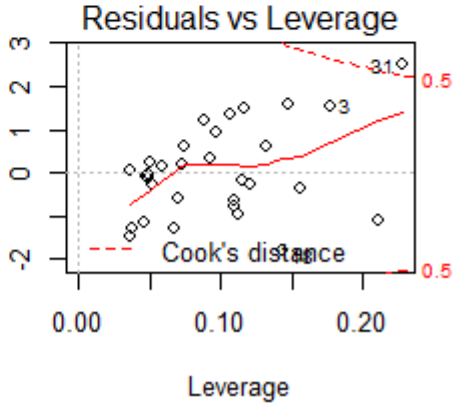
```
summary(fit)
par(mfrow=c(2,2))
plot(fit)
```

```
Call:
lm(formula = Volume ~ Girth + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```



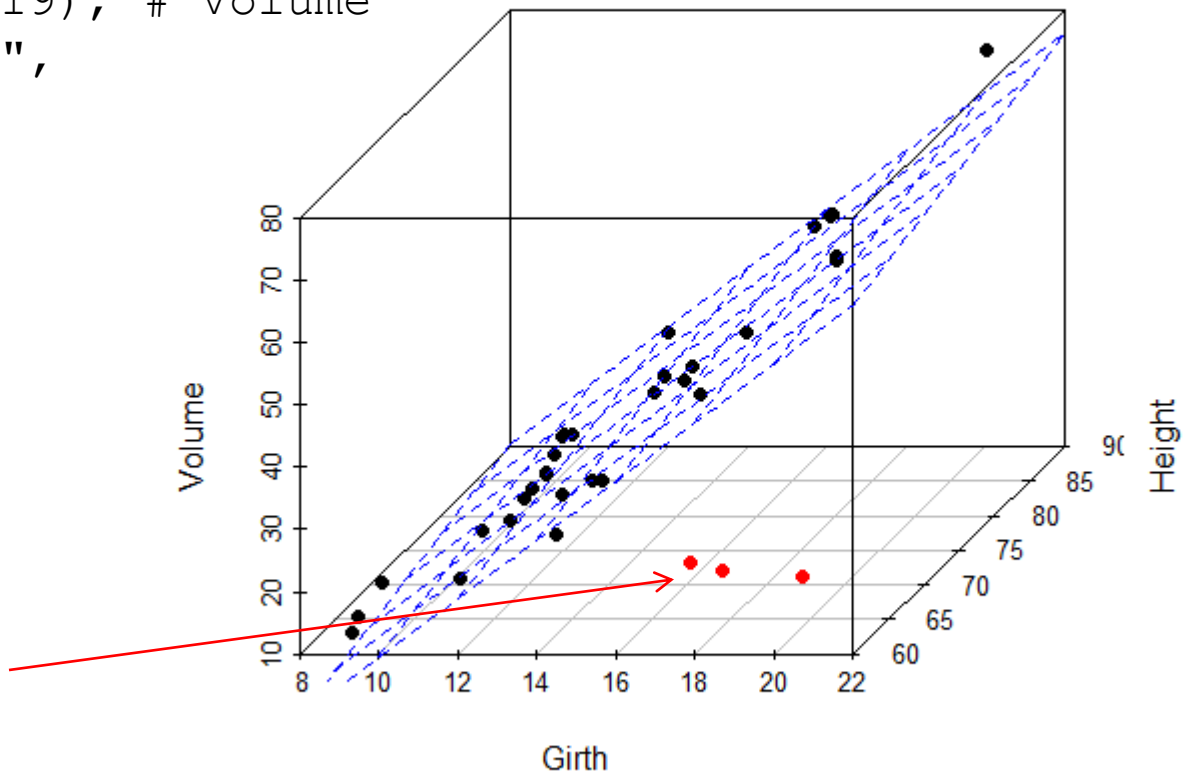


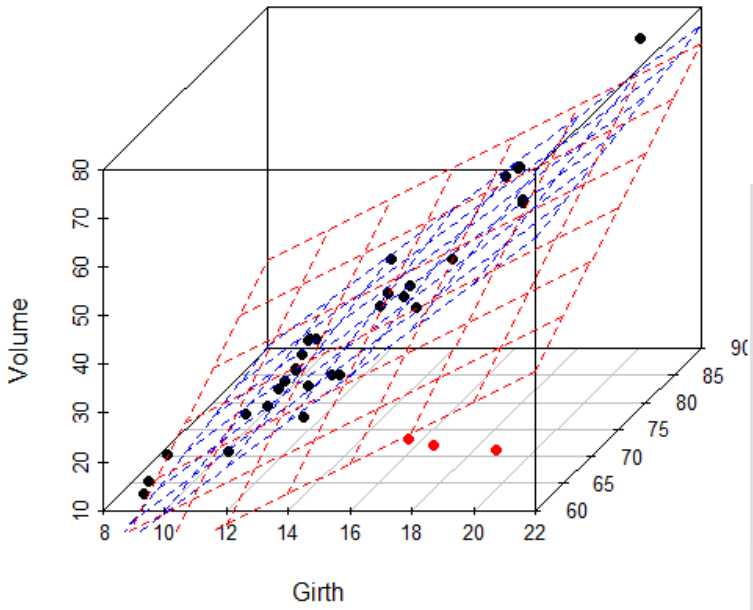




Volume ~ -57.9 + 4.7*Girth +0.33*Height

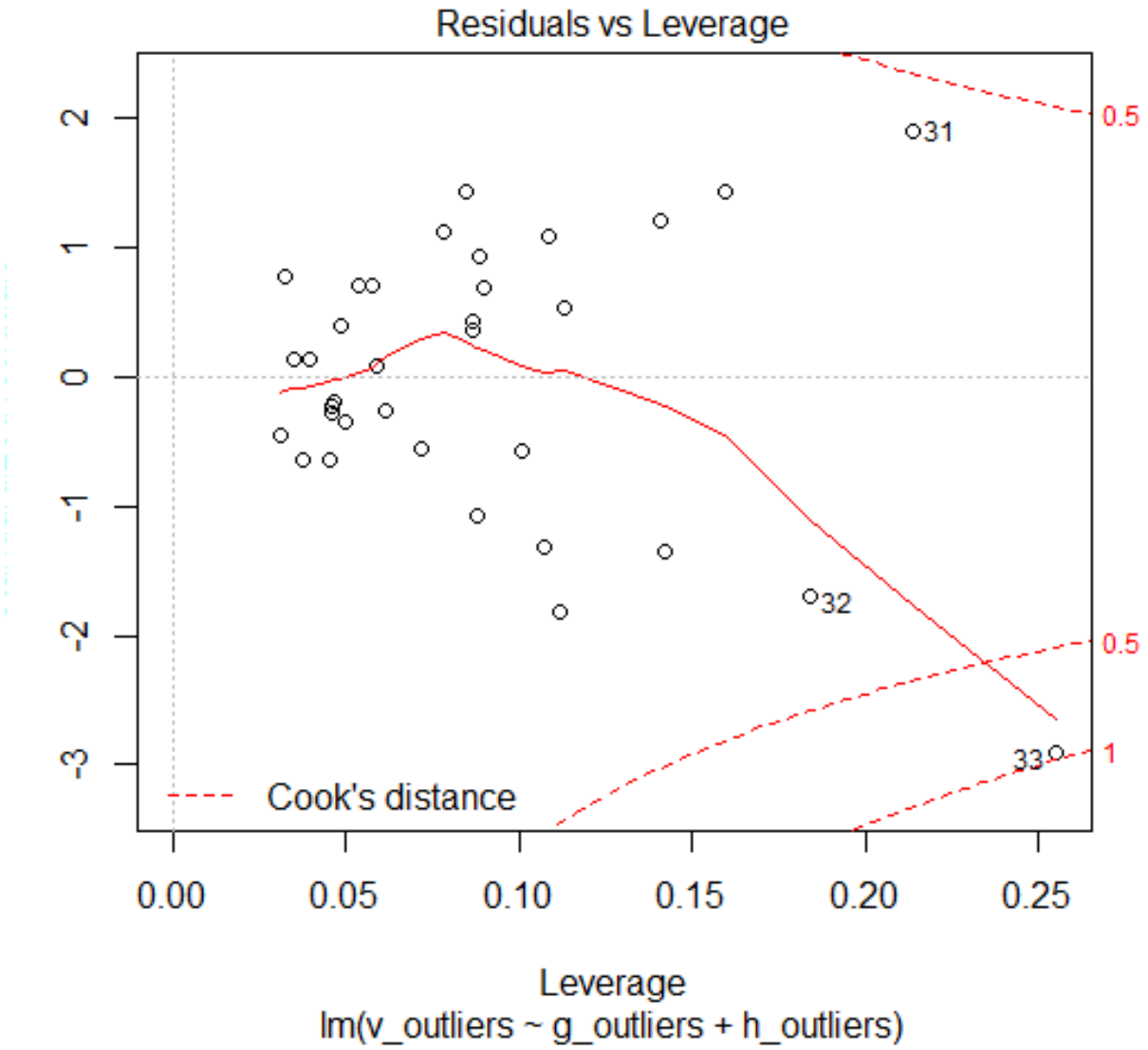
```
if(!require(scatterplot3d)){install.packages("scatterplot3d")}
graph<- scatterplot3d(Volume ~ Girth + Height,
                      pch=16,
                      angle=50)
# Plano regressao
graph$plane3d(fit, col="blue")
# Colocando Outliers
graph$points3d(c(18,20,17), # Girth
               c(64,64,65), # Height
               c(19,18,19), # Volume
               col="red",
               pch=16)
```



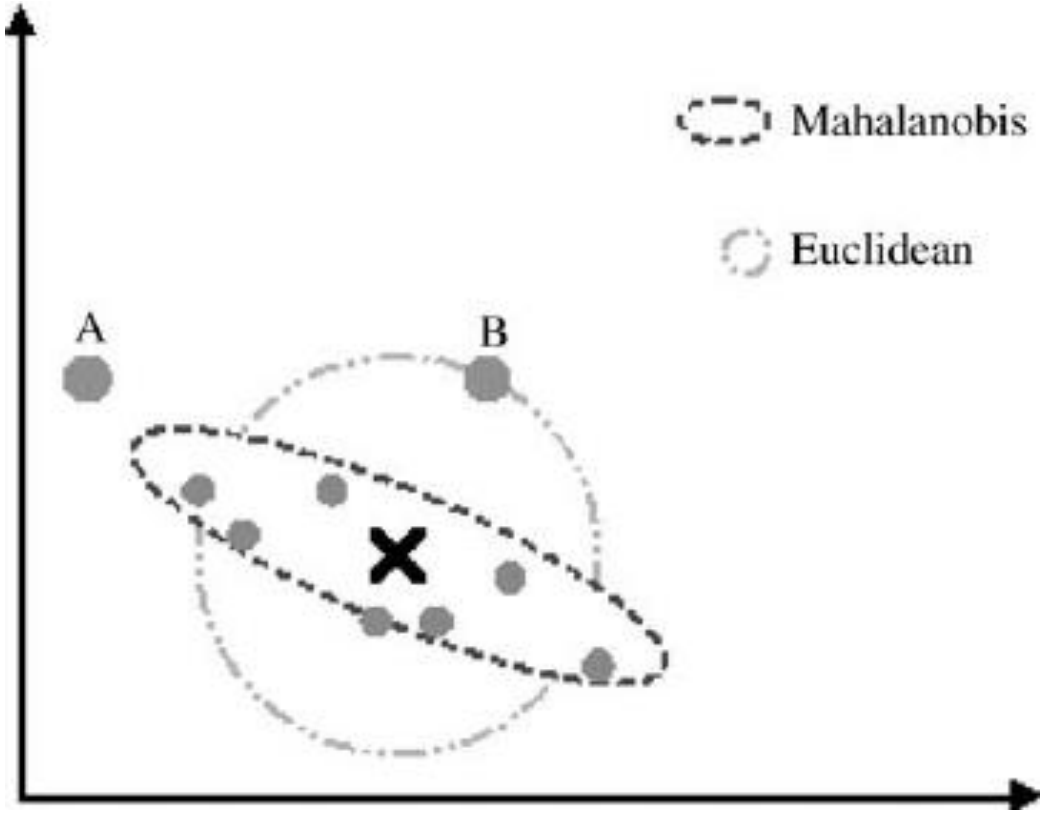



```
> g_outliers <- c(Girth,18,20,17)
> h_outliers <- c(Height,64,64,65)
> v_outliers <- c(Volume,19,18,19)
> fit2<-lm(v_outliers ~ g_outliers+h_outliers)
> graph$plane3d(fit, col="blue")
> graph$plane3d(fit2, col="red")
> summary(fit2)

Call:
lm(formula = v_outliers ~ g_outliers + h_outliers)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7523  -4.0505   0.8153   5.1864  12.6104

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -97.9567    14.4038  -6.801 1.29e-07 ***
g_outliers    3.1535     0.3969   7.946 5.70e-09 ***
h_outliers    1.1194     0.1916   5.842 1.92e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.499 on 31 degrees of freedom
Multiple R-squared:  0.7941,    Adjusted R-squared:  0.7808
F-statistic: 59.79 on 2 and 31 DF,  p-value: 2.294e-11
```

Distância de Cook



Distância de Mahalanobis



→Mede a distância entre um ponto **P** a uma distribuição **D**.

→Generalização da ideia de medir multidimensionalmente quantos desvios-padrões **P** está da média de **D**.

→É uma medida sem unidade

→Invariante na escala

→Leva em consideração as correlações entre os dados.

Fig: https://doi.org/10.1016/S1045-926X(03)00033-8

Pontos Influentes: DFFITS e DFBETA

Ver: http://www.portalaction.com.br/analise-de-regressao/343-pontos-influentes

$$DFFITS_{(i)} = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{QME_{(i)} h_{ii}}}, \qquad 2 \times \frac{\sqrt{(p+1)}}{(n-p-1)}$$

$$DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{QME_i c_{jj}}}, \qquad j = 0, 1, ..., p,$$

Para comparação entre:
→Cook's Distance
→Mahalanobis
→DFFITS
Ler: Oyeyemi *et al*. 2015

http://doi.org/10.5923/j.ajms.20150501.06

Distância de Cook

EXERCÍCIO

```
### Detectar Outliers na Regressao Multipla
trees_o <- cbind(g_outliers, h_outliers, v_outliers)
alfa <- 0.05          # probabilidade de erro tipo I para detectar outlier
p_val <- 1-alfa       # 1 - alfa
p <- ncol(trees_o)    # parametros utilizados
n <- nrow(trees_o)    # n observações

## Distancia de Cook
plot(fit2, which=5)
critico <- 0.2                              # definicao critica
abline(a=critico, b=0, col="red", lty=2)
cook <- as.vector(cooks.distance(fit2))
which(cook > critico)
```

Distância de Mahalonobis

EXERCÍCIO

```
# Distancia Mahalonobis
d <- mahalanobis(trees_o,            #  dados (p colunas)
                 colMeans(trees_o), # médias de cada p (centro de D)
                 cov(trees_o))      # matriz de covariancias (p x p)

plot(density(d, bw = 0.5),
     main=paste0("Distancias ao quadrado de Mahalanobis,
                  n= ",n, ",
                  p=",p))
rug(d)

gl <- p-1                                  # graus de liberdade neste caso
critico <-qchisq(p = p_val, df = gl)  # segue qui-quadrado
which(d > critico)
```

Pontos Influentes: DFFITS e DFBETA

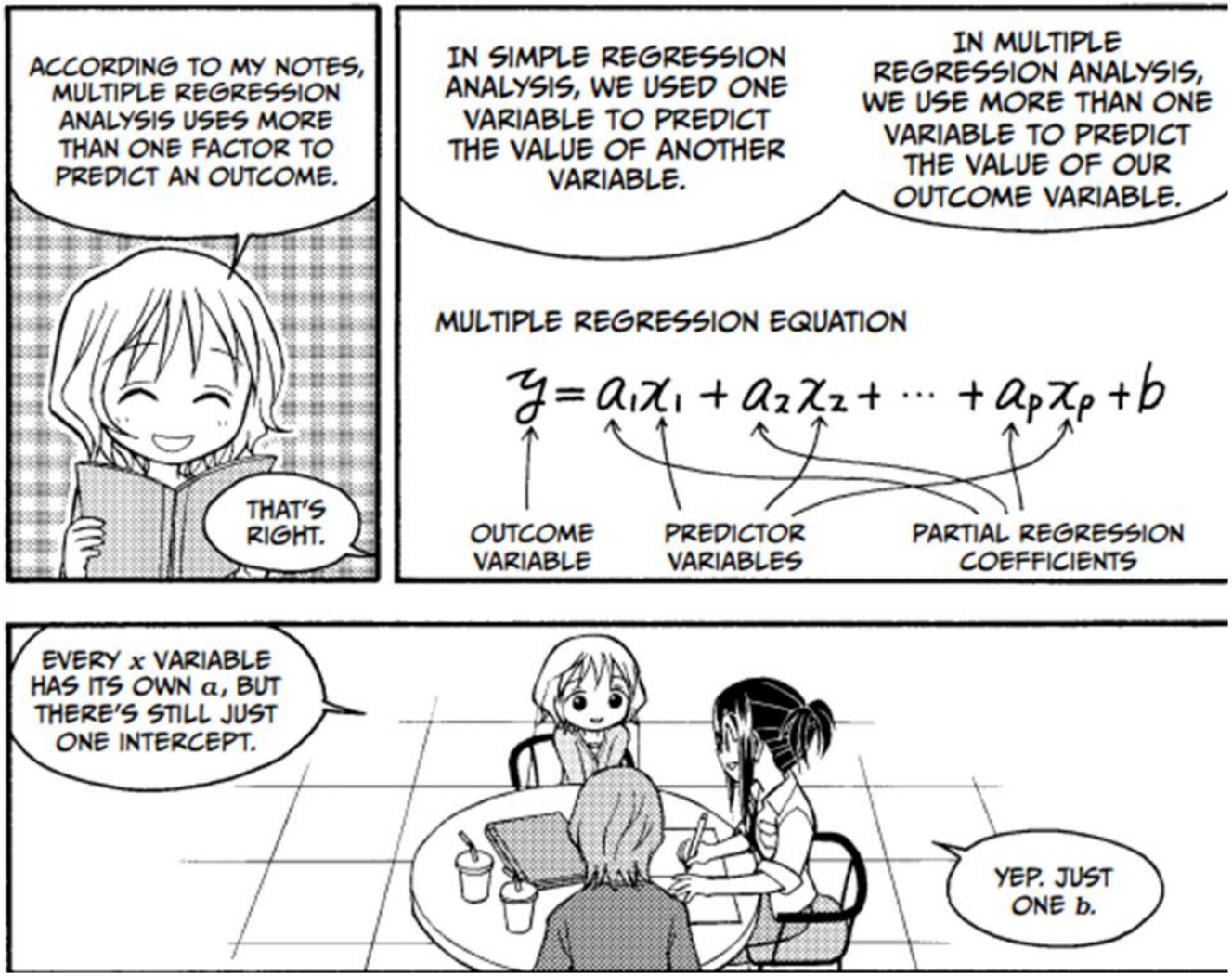
EXERCÍCIO

```
# Dffits - Desvio do Ajuste
ajuste <- as.vector(dffits(fit2))
critico <- 2 * (sqrt(p+1))/(n) # amostras maiores (>50)
critico <- 1                   # amostras pequenas (<50)
which(abs(ajuste) > critico)

# Dfbeta - desvio da inclinacao
d_beta <- as.vector(dfbeta(fit2))
critico <- 2/sqrt(n) # amostras grandes (n>50)
critico <- 1         # amostras pequenas (n<50)
which(abs(d_beta) > critico)
```

Resíduos "estudentizados"

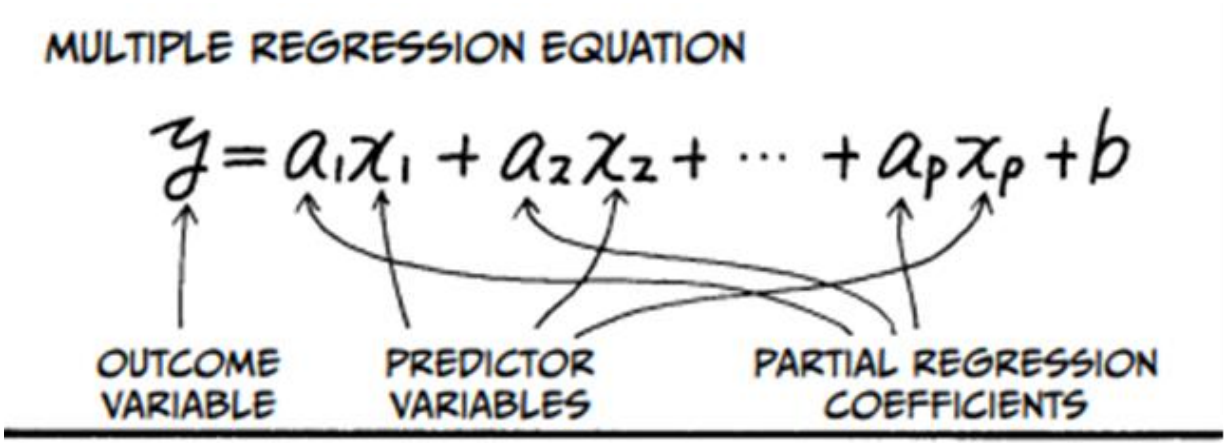
EXERCÍCIO

```
# Residuos estudentizados
require(MASS)
res_stu<-as.vector(studres(fit2))
plot(res_stu)
critico <- qt(p=p_val, df=n-1)
which(abs(res_stu) > critico)
```



Takashi & Inoue 2016.

The Manga Guide to Regression Analysis

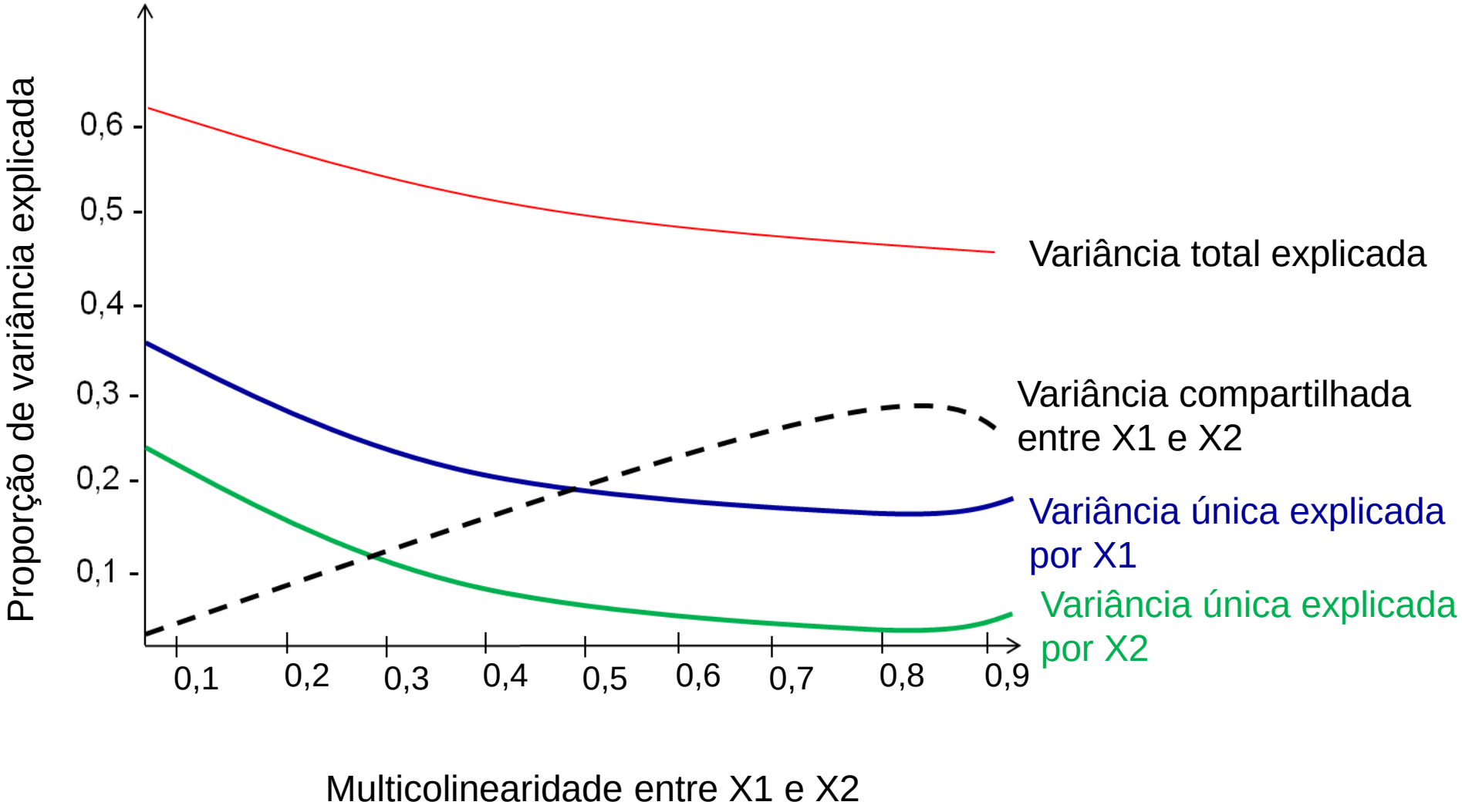


3 cuidados ao observarmos os valores de Beta:

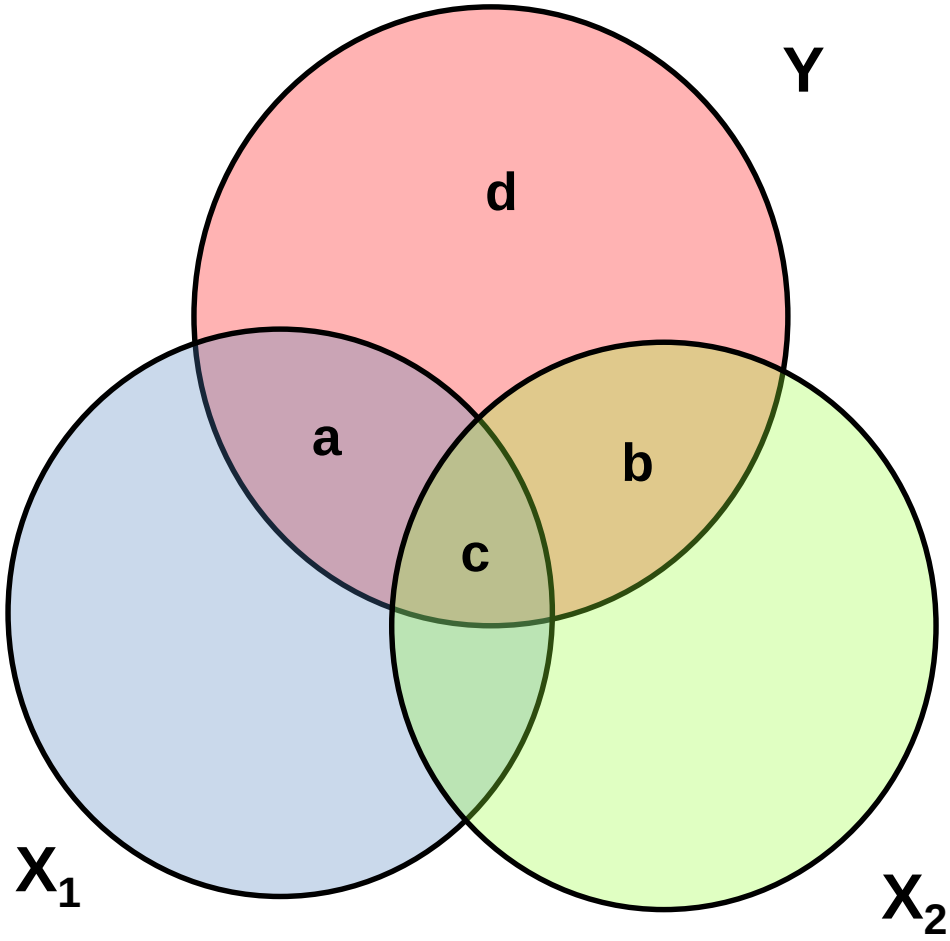
-Devem ser usados como uma orientação da importância relativa das variáveis independentes somente quando a MULTICOLINEARIDADE é mínima;

-Só podem sem interpretados no contexto das outras variáveis na equação;

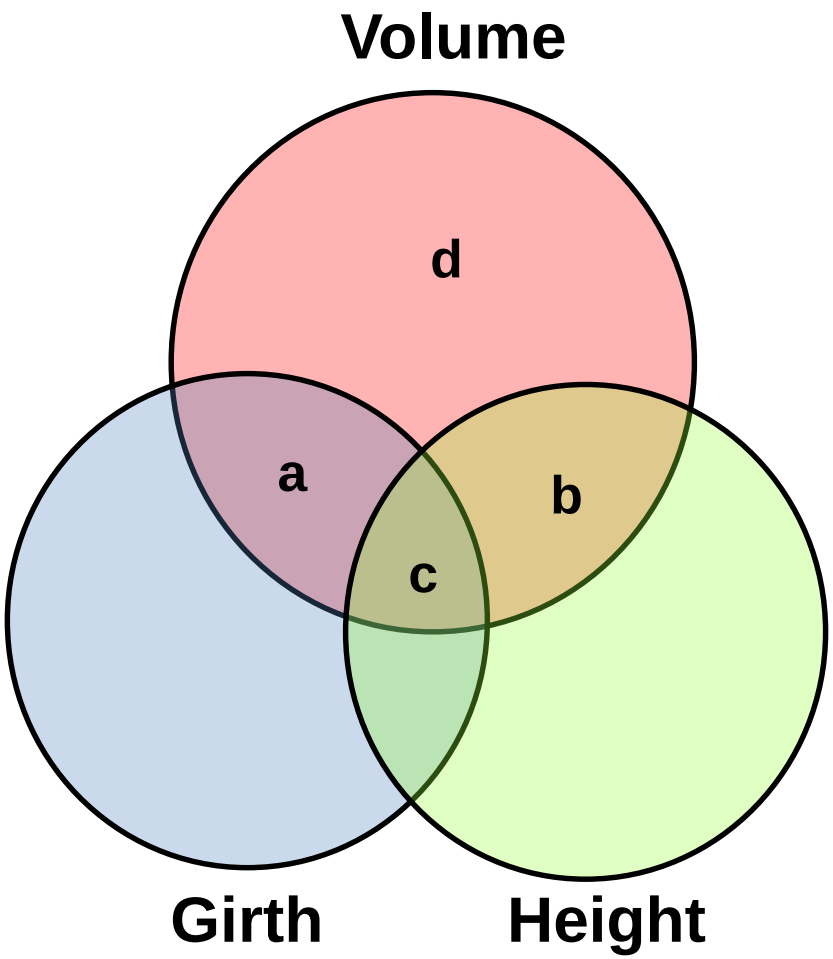
-Levam em consideração a escala e intervalo das variáveis independentes inseridas (transformação dos dados).



Baseado em: Hair et al. 2005. **Análise Multivariada de Dados.**



- **a**: Variância de Y unicamente explicada por X_1
- **b**: Variância de Y unicamente explicada por X_2
- **c**: variância de Y explicada conjuntamente por X_1 e X_2
- **d**: variância de Y não explicada por X_1 ou X_2



- **a** = rho_parcial_Girth²
- **b** = rho_parcial_Height²
- **c** =

  rho_parcial_Height² - rho_Height²

  ou

  rho_parcial_Girth² - rho_Girth²
- **d** = 1 – a – b – c

Correlação Parcial

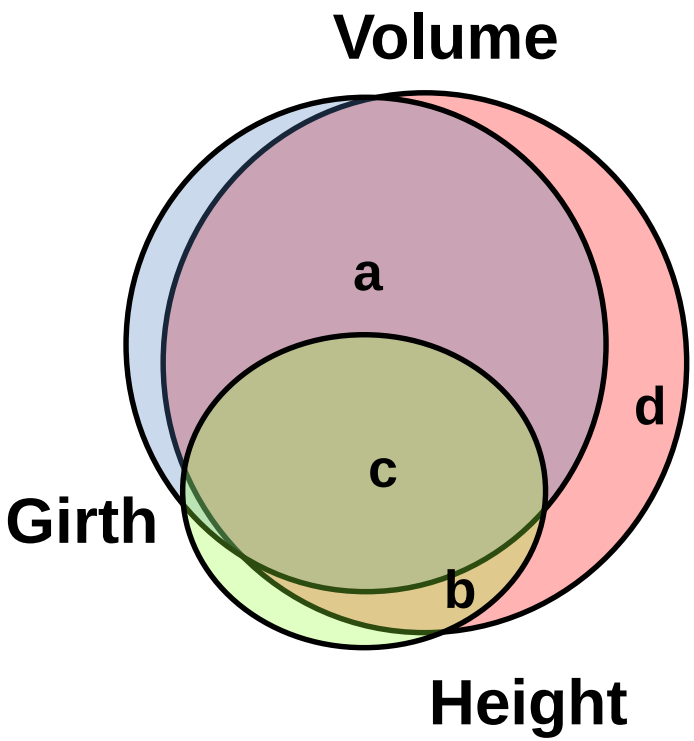
```
> round(cor(trees),2)
       Girth Height Volume
Girth   1.00   0.52   0.97
Height  0.52   1.00   0.60
Volume  0.97   0.60   1.00
```

Matriz de correlações

$$r_{Y,X_1(X_2)} = \frac{\rho_{Y,X_1} - \left(\rho_{Y,X_2} \times \rho_{X_1,X_2}\right)}{\sqrt{1{,}0 - \left(\rho_{X_1,X_2}\right)^2}}$$

```
rho_parcial <- function(Y=Y, X1=X1, X2=X2){
  rho_Y_X1 <- cor(Y, X1)
  rho_Y_X2 <- cor(Y, X2)
  rho_X1X2 <- cor(X1, X2)
  rho_parcial_X1 <- (rho_Y_X1 -
                     (rho_Y_X2*rho_X1X2))/
                       sqrt(1-rho_X1X2^2)
  return(rho_parcial_X1)
}
```

$$r_{Y,X_1(X_2)} = \frac{\rho_{Y,X_1} - \left(\rho_{Y,X_2} \times \rho_{X_1,X_2}\right)}{\sqrt{1,0 - \left(\rho_{X_1,X_2}\right)^2}}$$



- **a** = 0.590 → Só *Girth* já explica 59% da variância de *Volume*
- **b** = 0.013
- **c** = 0.345 → *Girth* + *Height* explica mais 34,5% da variância
- **d** = 0.052 → Não explicado pelo modelo

$$VIF = \frac{1}{Tolerância} = \frac{1}{1 - \rho_{parcial}}$$

Tolerância

Quantia de variabilidade da variável independente selecionada (X_n) **não é explicada** pelas outras variáveis independentes (X_m, sendo $n \neq m$).

→Valores baixos de Tolerância (altos de VIF) denotam alta colinearidade

VIF > 10 ou Tolerância < 0,10 são considerados limites, mas podem ser mais críticos (VIF > 5 ou Tolerância de 0,2, por exemplo)

VIF

1 = não correlacionado.

Entre 1 e 5 = moderadamente correlacionado

Maior que 5 = altamente correlacionado.

Problemas que a Multicolinearidade pode trazer:

1. **Altos valores de Erros padrões para os coeficientes;**
   Coeficiente geral pode se mostrar significativo, mas seus coeficientes individuais podem não ser significativos.
2. **Sinal incorreto de coeficientes;**
   Ao invés de apontar uma correlação positiva, acaba mostrando uma correlação negativa.
3. **Instabilidade**
   Quando as VI forem analisadas separadamente podem mostrar importâncias diferentes quando analisadas em conjunto.

Ações corretivas para a multicolinearidade

- Omitir uma ou mais variáveis independentes altamente correlacionadas (mas cuidado!);

- Usar somente como previsão (não para interpretar coeficientes de regressão);

- Usar correlações simples entre as variáveis independentes e dependente;

- Usar outras técnicas (PCA ou regressão Bayesiana).

+ Seleção de variáveis: *backward*, *forward* e *step-wise*. Ver:
https://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/
http://goodsciencebadscience.nl/?p=424

$$SS_{total} = \sum_i (y_i - \bar{y})^2$$

$$SS_{residuos} = \sum_i (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{residuos}}{SS_{total}}$$

$$\overline{R}^2 = 1 - \frac{SS_{residuos} / df_{erros}}{SS_{total} / df_{total}}$$

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1}$$

n: número de observações
k: número de variáveis independentes

EXERCÍCIO

```
require(datasets); require(car); data(swiss)
```

Interprete o modelo

```
fit <- lm(Fertility ~ ., data = swiss)
summary(fit)
```

Analise graficamente

```
par(mfrow=c(2,2)); plot(fit); par(mfrow=c(1,1))
```

Analise a correlação entre as variáveis. Há indícios de multicolinearidade?

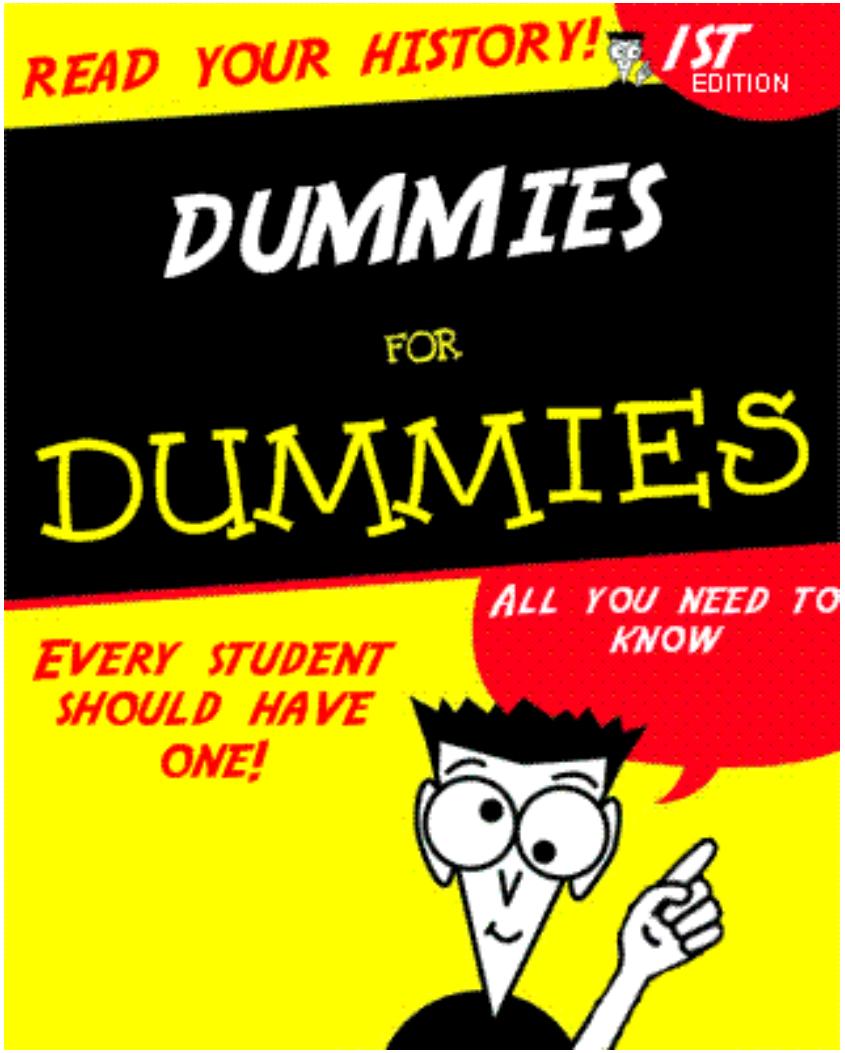
```
pairs(swiss, lower.panel=flines, upper.panel = fcor)
vif(fit)
```

Reescreva um novo modelo (com menos VI)

```
fit2 <- update(fit, . ~ Education + Catholic)
```

Analise os dois modelos por:

1. **ANOVA :** `anova(fit, fit2)`
2. **AIC:** `AIC(fit, fit2)`



Regressão Linear Múltipla

RLM COM VARIÁVEIS *DUMMY*

→Uma forma de incluir variáveis categóricas em modelos de regressão

→ **Variáveis indicadoras**: as variáveis categóricas no R devem ser vistas como uma variável `'factor'`.

→ A variável `'factor'` indica uma variável que possui **níveis** (`levels`) sendo, portanto, uma variável categórica típica dos modelos estatísticos.

→Utiliza transformação de variáveis em variáveis binárias

Exemplo:

Jogador	Posição
1	Armador
2	Armador
3	Ala
4	Ala
5	Pivô

Jogador	Armador	Ala	Pivô
1	1	0	0
2	1	0	0
3	0	1	0
4	0	1	0
5	0	0	1

```
egrandis <-
read.csv(http://ecologia.ib.usp.br/bie5782/lib/exe/fetch.php?media=dados:egrandis.csv, sep = ";")
# Tambem disponivel em "egrandis.csv"
summary(egrandis)
```

```
> mod <- lm( ht ~ dap + regiao, data=egrandis )
> summary(mod)

Call:
lm(formula = ht ~ dap + regiao, data = egrandis)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0361 -1.6546 -0.0863  1.5713  8.2935

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.79482    0.14107  19.811  < 2e-16 ***
dap            1.22698    0.01155 106.198  < 2e-16 ***
regiaoBotucatu 0.42383    0.15861   2.672  0.00759 **
regiaoItatinga -0.45400    0.14491  -3.133  0.00175 **
regiaoSalto    0.06980    0.12319   0.567  0.57104
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.315 on 2148 degrees of freedom
Multiple R-squared:  0.8432,    Adjusted R-squared:  0.8429
F-statistic:  2888 on 4 and 2148 DF,  p-value: < 2.2e-16
```

O modelo ajustado pela fórmula `'ht ~ dap + regiao'` é:

$$\mathrm{H}_i = \beta_0 + \beta_1 \mathrm{DAP}_i + \beta_2 I_{\mathrm{Botucatu}} + \beta_3 I_{\mathrm{Itatinga}} + \beta_4 I_{\mathrm{Salto}} + \varepsilon_i$$

onde:

$\mathbf{H}_i$: é a altura de cada árvore ***i*** ;

$\mathbf{DAP}_i$: é o Diâmetro a Altura do Peito (130 cm do chão) da árvore ***i*** ;

$I_{\mathbf{Botucatu}}$: é a **variável indicadora** para região de Botucatu, isto é, ela tem valor igual a `1` se `regiao` for igual a `Botucatu` e valor `0` (zero) se a `regiao` não for `Botucatu`;

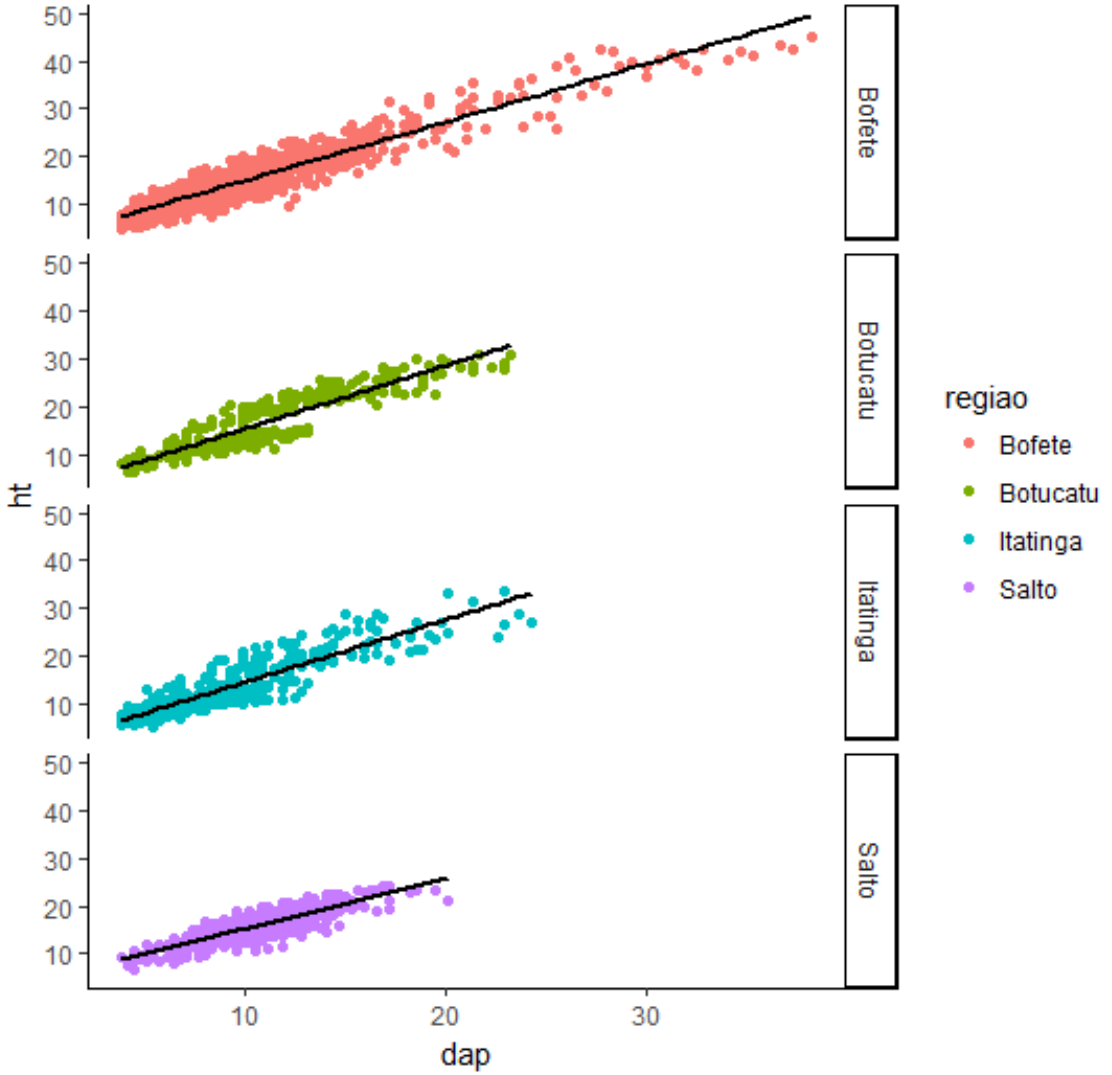
$I_{\mathbf{Itatinga}}$: é a **variável indicadora** para região de Itatinga.

E assim consecutivamente...

Logo, para **<u>Botucatu</u>** a predição será:

$$\mathrm{H}_i = \beta_0 + \beta_1 \mathrm{DAP}_i + \beta_2 1 + \beta_3 0 + \beta_4 0$$

```
if(!require(ggplot2)){install.packages("ggplot2")}
ggplot(egrandis, aes(x = dap, y = ht,
                     color = regiao, group = regiao)) +
  geom_point()+
  geom_smooth(method = "lm", se = F, color = "black") +
  facet_grid(regiao ~ .) +
  theme_classic()
```



```
model.matrix(mod)
amostra <-sample(1:nrow(egrandis), size=10)
model.matrix(mod)[amostra,]
```

Dados que são inseridos no modelo

```
> amostra <-sample(1:nrow(egrandis), size=10)
> model.matrix(mod)[amostra,]
     (Intercept)   dap regiaoBotucatu regiaoItatinga regiaoSalto
1576           1 12.10              0              0           1
464            1 10.50              0              0           1
650            1  4.46              0              0           0
19             1 10.19              0              0           1
213            1  4.46              0              0           1
1216           1 17.51              1              0           0
521            1  9.23              0              0           0
759            1  9.23              0              0           0
1938           1  4.14              0              0           0
590            1 14.01              0              0           0
>
```

Eucalipto de Botucatu

$$\mathrm{H}_i = \beta_0 + \beta_1 \mathrm{DAP}_i + \beta_2 I_{\mathrm{Botucatu}} + \beta_3 I_{\mathrm{Itatinga}} + \beta_4 I_{\mathrm{Salto}}$$

```
model.matrix(mod)
amostra <-sample(1:nrow(egrandis), size=10)
model.matrix(mod)[amostra,]
```

Dados que são inseridos no modelo

```
> amostra <-sample(1:nrow(egrandis), size=10)
> model.matrix(mod)[amostra,]
     (Intercept)   dap regiaoBotucatu regiaoItatinga regiaoSalto
1576           1 12.10              0              0           1
464            1 10.50              0              0           1
650            1  4.46              0              0           0
19             1 10.19              0              0           1
213            1  4.46              0              0           1
1216           1 17.51              1              0           0
521            1  9.23              0              0           0
759            1  9.23              0              0           0
1938           1  4.14              0              0           0
590            1 14.01              0              0           0
>
```

Eucalipto de Bofete

$$\mathrm{H}_i = \beta_0 + \beta_1 \mathrm{DAP}_i + \beta_2 I_{\mathrm{Botucatu}} + \beta_3 I_{\mathrm{Itatinga}} + \beta_4 I_{\mathrm{Salto}}$$

Modelo “Bruto”:

$$H_i = \beta_0 + \beta_1 DAP_i + \beta_2 I_{Botucatu} + \beta_3 I_{Itatinga} + \beta_4 I_{Salto} + \varepsilon_i$$

```
> mod <- lm( ht ~ dap + regiao, data=egrandis )
> summary(mod)

Call:
lm(formula = ht ~ dap + regiao, data = egrandis)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0361 -1.6546 -0.0863  1.5713  8.2935

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.79482    0.14107  19.811  < 2e-16 ***
dap            1.22698    0.01155 106.198  < 2e-16 ***
regiaoBotucatu 0.42383    0.15861   2.672  0.00759 **
regiaoItatinga -0.45400    0.14491  -3.133  0.00175 **
regiaoSalto    0.06980    0.12319   0.567  0.57104
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.315 on 2148 degrees of freedom
Multiple R-squared:  0.8432,    Adjusted R-squared:  0.8429
F-statistic:  2888 on 4 and 2148 DF,  p-value: < 2.2e-16
```

$$H_i = 2.79 + 1.22 \cdot DAP_i + 0.42 \cdot I_{Botucatu} - 0.45 \cdot I_{Itatinga} + 0.07 \cdot I_{Salto}$$

Interpretação

$$H_i = 2.79 + 1.22 \cdot DAP_i + 0.42 \cdot I_{Botucatu} - 0.45 \cdot I_{Itatinga} + 0.07 \cdot I_{Salto}$$

Botucatu tem um Intercepto 0.42 **maior** (2.79 + 0.42) do que Bofete

Correlação **positiva** entre DAP e H

Itatinga tem um Intercepto 0.46 **menor** (2.79 - 0.45) do que Bofete

É possível ajustar um **modelo de interação completo** do diâmetro com a variável região, alterando o *intercepto* **e** a *inclinação* do modelo em cada regiões:

```
> mod2 <- lm( ht ~ dap*regiao, data=egrandis )
> summary(mod2)

Call:
lm(formula = ht ~ dap * regiao, data = egrandis)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9287 -1.6110 -0.0882  1.5507  7.8854

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.87151    0.16558  17.342  < 2e-16 ***
dap                1.21961    0.01428  85.434  < 2e-16 ***
regiaoBotucatu    -0.72996    0.43535  -1.677 0.093747 .
regiaoItatinga    -1.26618    0.36273  -3.491 0.000491 ***
regiaoSalto        1.81343    0.43309   4.187 2.94e-05 ***
dap:regiaoBotucatu 0.10218    0.03620   2.823 0.004808 **
dap:regiaoItatinga 0.08382    0.03413   2.456 0.014123 *
dap:regiaoSalto   -0.16315    0.03906  -4.177 3.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.297 on 2145 degrees of freedom
Multiple R-squared:  0.8458,    Adjusted R-squared:  0.8453
F-statistic:  1681 on 7 and 2145 DF,  p-value: < 2.2e-16
```

Modelo de interação completo:

$$\mathrm{H}_i = \beta_0 + \beta_1 \mathrm{DAP}_i + \beta_2 I_{\text{Botucatu}} + \beta_3 I_{\text{Itatinga}} + \beta_4 I_{\text{Salto}} +$$

$$\beta_5 DAP_{\text{Botucatu}} + \beta_6 DAP_{\text{Itatinga}} + \beta_7 DAP_{\text{Salto}} + \varepsilon_i$$

```
> model.matrix(mod2)[amostra,]
     (Intercept)   dap regiaoBotucatu regiaoItatinga regiaoSalto dap:regiaoBotucatu dap:regiaoItatinga dap:regiaoSalto
1576           1 12.10              0              0           1              0.00                    0           12.10
464            1 10.50              0              0           1              0.00                    0           10.50
650            1  4.46              0              0           0              0.00                    0            0.00
19             1 10.19              0              0           1              0.00                    0           10.19
213            1  4.46              0              0           1              0.00                    0            4.46
1216           1 17.51              1              0           0             17.51                    0            0.00
521            1  9.23              0              0           0              0.00                    0            0.00
759            1  9.23              0              0           0              0.00                    0            0.00
1938           1  4.14              0              0           0              0.00                    0            0.00
590            1 14.01              0              0           0              0.00                    0            0.00
There were 15 warnings (use warnings() to see them)
>
```

Modelo de interação completo ou não?

```
> AIC(mod, mod2)
     df      AIC
mod   6 9731.324
mod2  9 9701.024
> anova(mod, mod2)
Analysis of Variance Table

Model 1: ht ~ dap + regiao
Model 2: ht ~ dap * regiao
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2148 11511
2   2145 11318  3    192.45 12.157 6.885e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Mais simples?

Parece que não...

```
> mod3<- lm( ht ~ dap, data=egrandis )
> mod4<- lm( ht ~ regiao, data=egrandis )
> AIC(mod, mod2, mod3, mod4)
     df       AIC
mod   6  9731.324
mod2  9  9701.024
mod3  3  9748.013
mod4  5 13675.047
> anova(mod, mod2, mod3, mod4)
Analysis of Variance Table

Model 1: ht ~ dap + regiao
Model 2: ht ~ dap * regiao
Model 3: ht ~ dap
Model 4: ht ~ regiao
  Res.Df   RSS Df Sum of Sq       F    Pr(>F)
1   2148 11511
2   2145 11319  3       192 12.1570 6.885e-08 ***
3   2151 11633 -6      -314  9.9304 7.798e-11 ***
4   2149 71950  2    -60317
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
> novo <- data.frame(dap=5, regiao="Botucatu")
> predict(mod2, novo, interval="confidence")
       fit      lwr      upr
1 8.750465 8.254166 9.246763
> predict(mod2, novo, interval="prediction")
       fit      lwr      upr
1 8.750465 4.218416 13.28251
```

Não inclui erro

Inclui erro

$\hat{H}_{novo}$

Ver: https://stats.stackexchange.com/questions/16493/difference-between-confidence-intervals-and-prediction-intervals

Regressão Linear Múltipla

TRATAMENTO E MODELAGEM PRÁTICA

EXERCÍCIO

Uma empresa pretende analisar o tempo médio do processo de atendimento observando o turno de entrada dos funcionários e o tempo de experiência deles. Os turnos de trabalho analisados foram manhã e tarde e o tempo de experiência é dado em dias.

Questões:

- Propor o modelo estatístico: Interpretação dos parâmetros do modelo, efeito das interações, suposições para o modelo;
- Estimação dos Parâmetros do Modelo;
- Análise de Variância;
- Medidas de associação: Coeficiente de determinação múltipla
- Testes Individuais e Intervalos de Confiança para os Parâmetros;
- Intervalo de Confiança para Resposta Média e Predição;
- Seleção de Variáveis;
- Seleção Todos os Modelos Possíveis;
- Análise de resíduos.

Fonte: http://www.portalaction.com.br/analise-de-regressao/exercicios

Dados: http://www.portalaction.com.br/sites/default/files/analise_regressao/planilhas/Exercicio%205%20-%20RegMult.xls (alterado) → **"Trabalho.csv" (sep=",", dec=".")**

EXERCÍCIO

Um banco pretende estudar a relação entre o volume de vendas de seguros efetuadas durante um dado período de tempo por seus vendedores, considerando seus anos de experiência e seu score num teste de inteligência.

Questões:

- Propor o modelo estatístico: Interpretação dos parâmetros do modelo, efeito das interações, suposições para o modelo;
- Estimação dos Parâmetros do Modelo;
- Análise de Variância;
- Medidas de associação: Coeficiente de determinação múltipla;
- Testes Individuais e Intervalos de Confiança para os Parâmetros;
- Intervalo de Confiança para Resposta Média e Predição;
- Seleção de Variáveis;
- Seleção Todos os Modelos Possíveis;
- Análise de resíduos.

Fonte: http://www.portalaction.com.br/analise-de-regressao/exercicios

Dados: http://www.portalaction.com.br/sites/default/files/analise_regressao/planilhas/Exercicio%203%20-%20RegMult.xls (alterado) → **"Vendas.csv" (sep=",", dec=".")**

Grato!


Google Scholar
ResearchGate
RPubs
% de Seriedade
% de me encontrar